Group-Sequential Strategies in Clinical Trials with Multiple Co-Primary Outcomes.

AUTHOR(S):

Hamasaki, Toshimitsu; Asakura, Koko; Evans, Scott R; Sugimoto, Tomoyuki; Sozu, Takashi

CITATION:

Hamasaki, Toshimitsu ...[et al]. Group-Sequential Strategies in Clinical Trials with Multiple Co-Primary Outcomes.. Statistics in biopharmaceutical research 2015, 7(1): 36-54

ISSUE DATE:

2015-03-18

URL:

http://hdl.handle.net/2433/198576

Group-sequential strategies in clinical trials with multiple co-primary outcomes

Toshimitsu Hamasaki[*,1,2], **Koko Asakura**[1,2], **Scott R Evans**[3], **Tomoyuki Sugimoto**[4], **and Takashi Sozu**[5]

[1] Office of Biostatistics and Data Management, National Cerebral and Cardiovascular Center, Japan
[2] Department of Innovative Clinical Trials and Data Science, Osaka University Graduate School of Medicine, Japan
[3] Department of Biostatistics and the Center for Biostatistics in AIDS Research, Harvard School of Public Health, USA
[4] Department of Mathematical Sciences, Hirosaki University Graduate School of Science and Technology, Japan
[5] Department of Biostatistics, Kyoto University School of Public Health, Japan

Final updated: June 26, 2015

We discuss the decision-making frameworks for clinical trials with multiple co-primary endpoints in a group-sequential setting. The decision-making frameworks can account for flexibilities such as a varying number of analyses, equally or unequally spaced increments of information and fixed or adaptive Type I error allocation among endpoints. The frameworks can provide efficiency, i.e., potentially fewer trial participants, than the fixed sample size designs. We investigate the operating characteristics of the decision-making frameworks and provide guidance on constructing efficient group-sequential strategies in clinical trials with multiple co-primary endpoints.

Key words: Adaptive Type I error allocation; Average sample number; equally or unequally spaced increments of information; Hierarchical testing procedure; Maximum sample size.

[*]Corresponding author: Toshimitsu Hamasaki, Office of Biostatistics and Data Management, National Cerebral and Cardiovascular Center, 5-7-1 Fujishirodai, Suita, Osaka 565-8565, Japan. e-mail: toshi.hamasaki@ncvc.go.jp

Group-sequential strategies in clinical trials with multiple co-primary outcomes

1 Introduction

Traditionally in clinical trials, one outcome is selected as a primary endpoint and used as the basis for the trial design including sample size determination, interim monitoring, and final analyses. However, as the clinical benefit of an intervention is often characterized by a set of (potentially correlated) multiple outcomes, many recent clinical trials, especially in medical product development, have utilized more than one endpoint as *co-primary* (Often et al., 2007). "Co-primary" in this setting means that the trial is designed to evaluate if the intervention is superior to the control on *all* of the endpoints. If the superiority for any of endpoints is not achieved, then the intervention fails to demonstrate the superiority to control. Note that, in contrast, designing the trial to evaluate an effect on at least one of the endpoints is a different problem, referred to as "multiple primary endpoints" or "alternative primary endpoints" (Often et al., 2007).

In complex diseases, co-primary endpoints may be preferable as they offer the opportunity of characterizing intervention's multidimensional effects. Regulators have issued guidelines recommending co-primary endpoints in several disease areas including Alzheimer's disease, acute heart failure, diabetes mellitus, Duchenne and Becker muscular dystrophy, and irritable bowel syndrome. For example, the Committee for Medicinal Products for Human Use (CMHP) issued a guideline recommending the use of cognitive, functional, and global endpoints to evaluate symptomatic improvement of dementia associated with Alzheimer's disease, indicating that primary endpoints should be stipulated reflecting the cognitive and functional disease aspects (CMHP, 2008). Offen et al. (2007) provides other examples with co-primary endpoints for regulatory purposes.

The resulting need for new approaches to the design and analysis of clinical trials with co-primary endpoints has been noted (Offen et al, 2007). Specifically controlling the Type I and Type II error rates when multiple K co-primary endpoints are potentially correlated is non-trivial. In hypothesis testing for the K co-primary endpoints, the null hypothesis is rejected if and only if all of the null hypotheses associated with each of the K endpoints are rejected at a significance level of α. No adjustment is needed to control the Type I error rate if each endpoint is tested at the same prespecified significance level. The corresponding rejection region of the null hypothesis, defined as the intersection of K regions associated with the K co-primary endpoints is considerable restricted and thus the hypothesis testing is conservative, especially when the number of endpoints to be evaluated is large. On the other hand, when designing the trial with K co-primary endpoints, the overall power should be maintained to evaluate the joint effects on all of the K endpoints. Since the Type II error rate increases as the number of endpoints increases, this requires the sample size adjustment and may often result in a sample size that is too large and impractical to conduct the clinical trial. In order to provide a more reasonable and practical sample size, methods for clinical trials with co-primary endpoints have been discussed in fixed sample size designs by many authors (Chuang-Stein et al., 2007; Hamasaki et al., 2013; Julious and Mclntyre, 2012; Kordzakhia et al., 2010; Offen et al, 2007; Senn and Bretz, 2007; Sozu et al., 2010, 2011, 2012, 2015; Sugimoto et al., 2012, 2013; Xiong et al., 2005). These methods commonly consider incorporating the correlations among the endpoints into the sample size calculation.

Hung and Wang (2009) discussed group-sequential strategies for clinical trials with multiple primary endpoints. These strategies provide the possibility of stopping a trial early when evidence is overwhelming, thus offering efficiency (i.e., potentially fewer patients than the fixed sample size designs). The methods also allow recalculation of the sample size based on the observed interim effects sizes. Recently Asakura et al. (2014, 2015) discuss two decision-making frameworks associated with hypothesis testing in clinical trials with two

continuous or binary endpoints as co-primary in a group-sequential setting. One framework is to reject the null hypothesis if and only if statistical significance is achieved for the two endpoints simultaneously (i.e., at the same interim timepoint of the trial). The other is a generalization of this, i.e., to reject the null hypothesis if superiority is demonstrated for the two endpoints at any interim timepoint (i.e., not necessarily simultaneously). The former framework is independently discussed by Chang et al. (2014) and evaluated in clinical trials with two co-primary endpoints. In the latter decision-making framework, Asakura et al. (2014, 2015) assume that the same number of analyses with a common information level between the two endpoints, and the Type I error allocation to each interim look should be specified and determined in advance, using any alpha-spending function method. However, the latter decision-making framework can be further generalized to accommodate a varying number of analyses and equally or unequally spaced increments of information among the endpoints.

In the decision-making framework above, the maximum Type I error rate associated with the rejection region of the null hypothesis for co-primary endpoints is not inflated over the prespecified significance level. However, the rejection region of the null hypothesis is still restricted similarly as in the fixed sample size designs, because there is a requirement that the allocation of Type I error to each interim analysis for all of the endpoints, be prespecified. To relax the rejection region of the null hypothesis for co-primary endpoints, the decision-making framework can be modified to allocate adaptively the Type I error to each interim look, using the methodology of hierarchical hypothesis testing with the adaptive Type I error allocation discussed in Tsong et al. (2004). However, Hung et al. (2007) cautions on the Type I error inflation in hierarchical hypothesis testing for detecting an effect on at least one endpoint in a group-sequential setting with multiple primary endpoints, and thus we need to investigate carefully how the Type I error rate behaves when using hierarchical hypothesis testing with the adaptive Type I error allocation in a group-sequential setting with multiple co-primary endpoints.

The flexibilities and extensions mentioned above may improve the power and rejection region of the tests, providing efficiency. However the decision-making and operational issues associated with the trial will be more complex and challenging. The objective of the paper is to investigate the operating characteristics (overall power, Type I error, and sample size) of the three decision-making frameworks for group-sequential strategies in clinical trials with multiple co-primary endpoints. The first two frameworks are the extensions of works in Asakura et al (2014) and Cheng et al (2014) to multiple co-primary endpoints when appropriately planning for a potentially varying number of analyses and information levels with the prespecified and fixed Type I error allocation. The last framework is an extension of the work in Tsong (2004) to multiple co-primary endpoints with adaptive Type I error allocation. We discuss the fundamental features of the three frameworks. We will not discuss methods for adaptation based on effects observed at interim of a trial. For sample size recalculation based on the conditional power, please see Asakura et al. (2014) and Cheng et al (2014). Asakura et al (2014) have extensively discussed and evaluated the sample size recalculation based on the conditional power with Cui-Hung-Wang statistics (Cui et al., 1999). This paper is structured as follows: in Section 2 we outline the decision-making frameworks for group-sequential strategies in clinical trials with multiple co-primary endpoints and briefly describe the power and sample size calculations in Section 3. In Section 4, we evaluate the operating characteristics of the three decision-making frameworks including power, Type I error rate and sample sizes. We summarize the findings and discuss advantages and disadvantages of the three decision-making frameworks in Section 5.

2 Group-sequential designs with co-primary endpoints

2.1 Statistical Settings

Consider a randomized, group-sequential clinical trial designed to compare test intervention (T) to control intervention (C), with K continuous outcomes being evaluated as co-primary endpoints ($K \geq 2$). Now suppose that a maximum of L analyses are planned. Let n_l and rn_l

be the cumulative number of participants on the test and the control intervention groups at the lth analysis ($l = 1, \dots, L$), respectively, where the sampling ratio ($r > 0$) is constant and not chosen arbitrarily during a clinical trial. Hence, up to n_L and rn_L participants are recruited and randomly assigned to either of the test and the control intervention groups, respectively. Then let responses to the test intervention denoted by $Y_{\mathrm{T}ki}$ and responses to the control intervention by $Y_{\mathrm{C}kj}$ $(k = 1, \dots, K;\ i = 1, \dots, rN; j = 1, \dots, (1-r)N)$. Assume that $(Y_{\mathrm{T}1i}, \dots, Y_{\mathrm{T}Ki})$ and $(Y_{\mathrm{C}1j}, \dots, Y_{\mathrm{C}Kj})$ are independently K-variate normally distributed as $(Y_{\mathrm{T}1i}, \dots, Y_{\mathrm{T}Ki}) \sim \mathrm{N}_K(\boldsymbol{\mu}_{\mathrm{T}}, \boldsymbol{\Sigma})$ and $\left(Y_{\mathrm{C}1j}, \dots, Y_{\mathrm{C}Kj}\right) \sim \mathrm{N}_K(\boldsymbol{\mu}_{\mathrm{C}}, \boldsymbol{\Sigma})$, respectively, where $\boldsymbol{\mu}_{\mathrm{T}}$ and $\boldsymbol{\mu}_{\mathrm{C}}$ are mean vectors given by $\boldsymbol{\mu}_{\mathrm{T}} = (\mu_{\mathrm{T}1}, \dots, \mu_{\mathrm{T}K})^{\mathrm{T}}$ and $\boldsymbol{\mu}_{\mathrm{C}} = (\mu_{\mathrm{C}1}, \dots, \mu_{\mathrm{C}K})^{\mathrm{T}}$ respectively. For simplicity, $\boldsymbol{\Sigma}$ is known covariance matrix given by $\boldsymbol{\Sigma} = \{\rho_{kk'}\sigma_k\sigma_{k'}\}$ with $\mathrm{var}[Y_{\mathrm{T}ki}] = \mathrm{var}[Y_{\mathrm{C}kj}] = \sigma_k^2$ and $\mathrm{corr}[Y_{\mathrm{T}ki}, Y_{\mathrm{T}k'i}] = \mathrm{corr}[Y_{\mathrm{C}kj}, Y_{\mathrm{C}k'j}] = \rho_{kk'} (k \neq k'; 1 < k < k' \leq K; K \geq 2)$.

Let δ_k denote the differences in the means for the test and the control intervention groups respectively, where $\delta_k = \mu_{\mathrm{T}k} - \mu_{\mathrm{C}k} (k = 1, \dots, K)$. Suppose that positive values of δ_k represent the test intervention's benefit. We are interested in testing the null hypothesis $\mathrm{H}_0: \delta_k \leq 0$ for at least one k versus the alternative hypothesis $\mathrm{H}_1: \delta_k > 0$ for all k. Let $(Z_{1l}, \dots, Z_{Kl})$ be the statistics for testing the hypotheses at the lth analysis, given by $Z_{kl} = (\bar{Y}_{\mathrm{T}kl} - \bar{Y}_{\mathrm{C}kl})/(\sigma_k\sqrt{(1+r_l)/(n_l r_l)})$ where $\bar{Y}_{\mathrm{T}kl}$ and $\bar{Y}_{\mathrm{C}kl}$ are the sample means given by $\bar{Y}_{\mathrm{T}kl} = n_l^{-1}\sum_{i=1}^{n_l} Y_{\mathrm{T}ki}$ and $\bar{Y}_{\mathrm{C}kl} = (r_l n_l)^{-1}\sum_{j=1}^{r_l n_l} Y_{\mathrm{C}kj}$. Thus, each Z_{kl} is normally distributed as $\mathrm{N}(\sqrt{r_l n_l/(1+r_l)}\delta_k/\sigma_k, 1^2)$ under H_1. As the joint distribution of $(Z_{1l}, \dots, Z_{Kl})$ is K-variate normal with the correlation $\rho_{kk'}$ and the joint distribution of $(Z_{k1}, \dots, Z_{kL})$ is L-variate normal with the correlation $\sqrt{n_l/n_{l'}}$ $(1 \leq l \leq l' \leq L)$, the joint distribution of $(Z_{1l}, \dots, Z_{Kl}, \dots, Z_{1L}, \dots, Z_{KL})$ is $K \times L$ multivariate normal with their correlation given by $\rho_{kk'}\sqrt{n_l/n_{l'}}$ $(k \neq k'; l \neq l')$.

2.2 Decision-making framework A: Prespecified and fixed Type I error allocation

When evaluating the joint effects on all K endpoints within the context of group-sequential designs, a general decision-making framework associated with hypothesis testing is to reject H_0 if statistical significance of a test intervention relative to control is achieved for all endpoints at any interim timepoint until the final analysis (i.e., not necessarily simultaneously) (DF-A). If superiority is demonstrated on some but not all of the endpoints at the interim, then the trial will continue but subsequent hypothesis testing is repeatedly conducted only for the previously non-significant endpoint(s). Thus DF-A offers the opportunity of stopping measurement of an endpoint for which superiority has already been demonstrated. Stopping measurement may be desirable if the endpoint is very invasive or expensive (e.g., data from a liver biopsy or gastro-fiberscope, or data from expensive imaging). In addition, DF-A is a flexible strategy that allows the option of selecting different timings for interim looks among the endpoints. For example, when two endpoints are considered as co-primary and the number of analyses is four for one endpoint and three for the other endpoint, DF-A can allow for information times of 0.25, 0.50, 0.75 and 1.0 for one endpoint and 0.33, 0.67 and 1.0 for the other endpoint. However, the different timings for interim looks may create operational difficulty in conducting a clinical trial. For practical purposes, in Section 4, we will consider a situation where the timing of interim looks is the same among the endpoints, e.g., 0.25, 0.50, 0.75 and 1.0 for one endpoint and 0.50, 0.75 and 1.0 for the other endpoint.

Here suppose that L_k analyses are planned for each endpoint and a total number of analyses L is the sum of the number of analyses over all of the endpoints excluding the duplications of the same information time $n_{l_k}/n_L = n_{l_{k'}}/n_L$. The stopping rule for DF-A is formally given as follows:

Until the lth analysis ($l = 1, \dots, L-1$),

If $Z_{kl_k} > c_{kl_k}$ for all K endpoints for some $1 \le l_k \le l$, then reject H_0 and stop the trial,

otherwise, continue to the $(l+1)$th analysis,

at the Lth analysis,

if $Z_{kL_k} > c_{kL_k}$ for non-significant endpoint(s) until the $(L-1)$th analysis, then reject H_0,

otherwise, do not reject H_0,

where Z_{kl_k} are the test statistics at the l_kth analysis for the kth endpoint, c_{kl_k} are the critical values at the l_kth analysis for the kth endpoint. Note that c_{kl_k} are prespecified and determined separately, using any group-sequential methods such as the Lan-DeMets (LD) alpha-spending method (Lan and DeMets, 1984) to control an overall Type I error rate of α, as if they were a single primary endpoint, ignoring the other co-primary endpoint(s). Therefore, the overall power (or conjunctive power) corresponding to DF-A is

$$1-\beta = \Pr\left[\left\{\cup_{l_1=1}^{L_1}\{Z_{1l_1} > c_{1l_1}\}\right\} \cap \cdots \cap \left\{\cup_{l_K=1}^{L_K}\{Z_{Kl_K} > c_{Kl_K}\}\right\} | H_1\right]. \quad (1)$$

DF-A is flexible, but stopping measurement may also introduce operational challenges into the trial. To avoid the operational difficulties, we may opt for a restriction regarding when H_0 is rejected and the trial is stopped. The simplified version of DF-A is to reject H_0 if superiority is demonstrated on all of the endpoints at an interim simultaneously. If any of the endpoints is not significant, then then the trial continues until the joint significance for all endpoints is established simultaneously (DF-A'). The stopping rule for DF-A' is formally given as follows:

At the lth analysis ($l = 1, \dots, L$),

If $Z_{kl} > c_{kl}$ for all K endpoints simultaneously, then reject H_0 and stop the trial,

otherwise, continue to the $(l+1)$th analysis,

at the Lth analysis,

if $Z_{kL} > c_{kL}$ for all K endpoints simultaneously, then reject H_0,

otherwise, do not reject H_0.

Therefore, the overall power corresponding to DF-A' is a special case of DF-A,

$$1 - \beta = \Pr\left[\cup_{l=1}^{L}\{\{Z_{1l} > c_{1l}\} \cap \cdots \cap \{Z_{Kl} > c_{Kl}\}\} | H_1\right]. \quad (2)$$

DF-A' is simpler but less powerful than DF-A. This will be illustrated in Section 4.

2.3 Decision-making framework B: Hierarchical hypothesis testing with adaptive Type I error allocation

For the methods discussed in the previous section, the rejection region of the null hypothesis is still restricted, as with the fixed sample size designs, because the allocation of Type I error to each interim analysis for all endpoints should be prespecified using an alpha-spending method. To overcome the issue, the decision-making framework can be modified to allocate adaptively the Type I error to each interim look, using the methodology of hierarchical hypothesis testing with adaptive Type I error allocation. This idea is discussed by Tsong et al. (2004) in group-sequential three-arm clinical trials when assessing the equivalence and efficacy of a generic product, where the co-primary objectives of the study are to assess whether the generic and reference product are effective relative to placebo and whether the generic is equivalent to the reference product with a prespecified equivalence margin. Their method evaluates equivalence only after both null hypotheses of efficacy are rejected and then to specify the Type I error allocation before the equivalence evaluation is performed.

When extending the hierarchical hypothesis testing with adaptive Type I error allocation to clinical trials with multiple endpoints as co-primary, the order of the hypothesis testing for each endpoint is determined even when the endpoints are equally important and the Type I error allocation for the first-tested endpoint is prespecified, using an alpha spending method, where a maximum of planned analyses for the first-tested endpoint is L_1. If superiority is

established for the first-tested endpoint at l_1 th analysis with information time $I_{l1} = n_{l1}/n_L\,(0 < I_{l1} \leq 1)$, then the Type I error allocation for the second-tested endpoint is specified before the hypothesis testing for the second-tested endpoint is performed, where a maximum of planned analyses for the second-tested endpoint is L_2. If superiority has been established for the second-tested endpoint at l_2th analysis with information time $I_{l2} = n_{l2}/n_L$ $(I_{l1} \leq I_{l2} \leq 1)$, then the Type I error allocation for the third-tested endpoint is specified before the hypothesis testing for the third-tested endpoint is performed. These steps are repeated for Kth-tested endpoint until H_0 is rejected. The stopping rule for DF-B is formally given as follows:

For kth-tested endpoint $(1 \leq k \leq K)$, at the l_kth analysis $(l_k = 1, \ldots, L_k - 1)$,

If $Z_{kl_k} > c_{kl_k}$, then specify the Type I error allocation for $(k+1)$th-tested endpoint,

using any alpha-spending method

otherwise, continue to the $(l_k + 1)$th analysis,

at the L_kth analysis,

if $Z_{kL_k} > c_{kL_k}$, then specify the Type I error allocation for $(k+1)$th-tested endpoint,

using alpha-spending methods

otherwise, do not reject H_0.

For Kth-tested endpoint at the l_Kth analysis $(l_K = 1, \ldots, L_K - 1)$,

if $Z_{Kl_K} > c_{Kl_K}$, then reject H_0 and stop the trial,

otherwise, continue to the $(l_K + 1)$th analysis,

at the L_Kth analysis,

if $Z_{KL_K} > c_{KL_K}$, then reject H_0 and stop the trial,

otherwise, do not reject H_0.

For example, consider a clinical trial with two co-primary endpoints, where the maximum number of analyses for the first-tested endpoint is L_1 =5, with equally spaced increments of

information and the O'Brien-Fleming-type boundary is used to reject the null hypothesis for the first-tested endpoint with the significance level of α =2.5% for a one-sided test. The second-tested endpoint is evaluated only after the null hypothesis for the first-tested endpoint is rejected. The second endpoint is tested at the remaining planned analyses for the first-tested endpoint, and the O'Brien-Fleming-type boundary (O'Brien and Fleming, 1979) is used to reject the null hypothesis for the second-tested endpoint with the significance level of α =2.5% for a one-sided test, as shown in Table 1. If the first-tested endpoint is statistically significant at the 4th look, then the second-tested endpoint is tested twice with the boundary of 2.2504 at 4th analysis and 2.0249 at the final analysis.

The overall power for DF-B is

$$1-\beta = \Pr\left[\cup_{l_1=1}^{L_1}\left\{\{Z_{1l_1} > c_{1l_1}\} \cap \left\{\cdots \cap \left\{\cup_{l_K=1}^{L_K}\{Z_{Kl_K} > c_{Kl_K}\}\right\}\right\}\right\} | \mathrm{H}_1\right]. \quad (3)$$

For the sample size calculation, the number of interim analyses and the information time for all of the endpoints should be prespecified. As mentioned in Section 1, Hung et al. (2007) discuss the behavior of the Type I error rate when hierarchical hypothesis testing is used for detecting an effect on *at least one* endpoint in a group-sequential setting and caution that the conventional hierarchical testing strategy may violate the closed testing principle and thus the overall Type I error rate may not be controlled in the strong sense. They show that, when considering the two endpoints as primary and testing the two hypotheses for the two endpoints with the hierarchical order, the Type I error rate for the second endpoint is inflated over the prespecified significance level, depending on the effect size for the first endpoint and correlation between the endpoint. Thus DF-B may not control the Type I error rate adequately. This will be further evaluated in Section 4 and the Appendix.

3. Calculation for power and sample sizes

The powers (1), (2) and (3) defined in the previous sections can be evaluated using the numerical integration method in Genz (1992) or other methods. The power calculation

requires considerable computing time and memory especially with a large number of endpoints or number of analyses. The accuracy of the computation should be carefully controlled as it is sensitive to the number of endpoints and the number of analyses.

We describe two sample size concepts, i.e., the maximum sample size (MSS) and the average sample number (ASN) (i.e., expected sample size) based on the power (1), (2) or (3). The MSS is the sample size required for the final analysis to achieve the desired power $1-\beta$. The MSS is given by the smallest integer not less than n_L satisfying the power for a group-sequential strategy at the prespecified δ_k and $\rho_{kk'}$, with Fisher's information time for the interim analyses, n_l/n_L $(l=1,\ldots,L)$. To identify the value of n_L, an easy strategy is a grid search to gradually increase (or disease) n_L until the power under n_L exceeds (or falls below) the desired power. As seen in Appendix 1, the grid search often requires considerable computing time, especially with a larger number of endpoints, a larger number of analyses, or a small effect size. To reduce the computing time, the Newton–Raphson algorithm in Sugimoto et al. (2012) or the basic linear interpolation algorithm in Hamasaki et al. (2013) may be utilized. In this paper, we use of the basic linear interpolation algorithm to reduce the computing time.

The ASN is the expected sample size under hypothetical reference values and provides information regarding the number of participants anticipated in a group-sequential design in order to reach a decision point, and the ASN per intervention group is given by

$$\mathrm{ASN}=\sum_{l=1}^{L-1} n_l P_l(\delta_1,\ldots,\delta_K)+n_L\left(1-\sum_{l=1}^{L-1} P_l(\delta_1,\ldots,\delta_K)\right),$$

where $P_l(\delta_1,\ldots,\delta_K)$ is stopping probability (or exit probability) as defined the likelihood of crossing the critical boundaries at the lth interim look assuming the true values of the intervention's effect are $(\delta_1,\ldots,\delta_K)$.

Both MSS and ASN depend on the design parameters including differences in means, the correlation structure among the endpoints, the selected stopping boundary based on LD alpha-

spending method (e.g., O'Brien-Fleming-type boundary, Pocock-type boundary (Pocock, 1977)), the number of analyses, and equally or unequally spaced increments of information.

Our experience suggests that, as shown in Appendix, when considering more than two endpoints as co-primary in a group-sequential setting with more than five analyses, calculating the multivariate normal integrals often requires considerable computing time. A Monte-Carlo simulation-based method provides an alternative but the number of replications for simulations should be carefully chosen to control simulation error in calculating the empirical power.

4. Operating characteristics of the decision-making frameworks in group-sequential strategies

In this section, we investigate the operating characteristics of the decision-making frameworks for the group-sequential strategies described in the previous section including the overall Type I error rate, overall power and ASN under a given sample size, of one-sided test. and For illustration, we consider a simple situation, i.e., a randomized clinical trial designed to compare a test intervention to a control intervention with two outcomes being evaluated as co-primary endpoints. We evaluate the operating characteristics of the decision-making frameworks for group-sequential strategies shown in Tables 2 and 3. They include clinical trials with a maximum number of analyses of 2 or 5, and equally spaced increments of information for one endpoint, but unequally spaced increments for other endpoint, with a common variance $\sigma_1 = \sigma_2 = 1.0$. One-sided statistical testing is conducted at the significance level of $\alpha = 2.5\%$. A range of correlation between the two outcomes considered in the evaluation is , $\rho_{12} \geq 0$ since the correlation among the endpoints are usually nonnagetive as discussed in Often et al (2007). The overall power and Type I error rate is evaluated using the numerical integration method in Genz (1992). However, the accuracy of the computation for the overall power and Type I error rate may depend on the number of analyses. Therefore,

Monte-Carlo simulation was also performed to confirm the result from the numerical integration method. A total of 100,000 replications and 1,000,000 replications are selected for the assessments of power and Type I error rate respectively. The number of replications was determined based on the precision, where a sample size of 1,000,000 provides a two-sided 95% confidence interval with a width equal to 0.001 when the proportion is 0.025, and 100,000 replications provides a two-sided 95% confidence interval with a width equal to 0.005 when the proportion is 0.80. The results presented in this manuscript were by the numerical integration methods, but the Monte-Carlo simulation confirmed these results.

4.1 Behaviors of the overall Type I error rate

Figures 1 (L =2) and 2 (L =5) illustrate the behaviors of the Type I error rate with correlation under a given sample size per group (equally-sized groups: r =1) in the decision-making frameworks for group-sequential strategies with two co-primary endpoints as shown in Tables 2 and 3. The effect size (δ_1, δ_2) selected were (0.2, 0.2), (0.3, 0.2) and (0.2, 0.3) and the given sample sizes per group are calculated to detect the joint effect of (δ_1, δ_2) with the power of 80% at the significance level of 2.5% for a one-sided test in a fixed sample size design; they are 516 for $(\delta_1, \delta_2) = (0.2,\ 0.2)$, and 402 for $(\delta_1, \delta_2) = (0.3,\ 0.2)$ or $(\delta_1, \delta_2) = (0.2,\ 0.3)$, which The critical values are determined based on the O'Brien–Fleming-type boundary (OF) (O'Brien and Fleming, 1979), Pocock-type boundary (PC) (Pocock, 1979) or their combinations, using with the LD alpha-spending method. Then four stopping boundary combinations are considered: (i) the OF for both endpoints (OF-OF), (ii) the OF for δ_1 and the PC for δ_2 (OF-PC), (iii) the PC for δ_1 and the OF for δ_2 (PC-OF), and (iv) the PC for both endpoints (PC-PC). The overall Type I error rate is evaluated under three pairs of the mean differences $(\delta_1, \delta_2) = (0.0, 0.0)$, (0.0, 0.2) and (0.2, 0.0).

In the case of L =2, in all stopping boundary combinations and effect size combinations, Strategy #2 (DF-A') is the most conservative as it provides the smallest Type I error rate

among the strategies. For Strategies #2 (DF-A'), #3 (DF-A) and #4 (DF-A), the Type I error rate increases as the correlation goes toward one, but does not exceed the targeted significance level of 2.5% in all of the stopping boundary combinations and effect size combinations. Strategy #4 provides a larger Type I error rate than Strategy #3, illustrating that delaying the analysis for Endpoint 2 relaxes the Type I error rate.

For Strategy #1 (DF-B), similarly as seen in Strategies #2, #3 and #4, the Type I error rate increases as the correlation goes toward one, but does not exceed the targeted significance level of 2.5%, in all of the stopping boundary combinations and effect size combinations except for $(\delta_1, \delta_2) = (0.2,\ 0.0)$. However, in all of the stopping boundary combinations with effect size combination $(\delta_1, \delta_2) = (0.2,\ 0.0)$, it does exceed the targeted significance level of 2.5%, especially in the stopping boundary combination of PC-OF or PC-PC.

In the case of $L = 5$, the behaviors of the Type I error rate are similar to that seen with $L = 2$. Strategy # 2 (DF-A') is the most conservative as it provides the smallest Type I error rate among the strategies. For Strategies #3 to #6, the Type I error rate increases as the correlation goes toward one, but does not exceed the targeted significance level of 2.5% in all of the stopping boundary and effect size combinations. However, for Strategy #1, in all of the stopping boundary combinations with effect size combination $(\delta_1, \delta_2) = (0.2,\ 0.0)$, it does exceed the targeted significance level of 2.5%.

The two decision-making frameworks with the prefixed Type I error allocation adequately controls the Type I error rate but the decision-making framework with the adaptive Type I error allocation may not control the Type I error rate. The Type I error rate is inflated depending on the correlation, effect sizes, and stopping boundary. Details are provided in the Appendix. Further investigation is needed to understand how the Type I error rate for the DF-B behaves in the original context of Tsong et al. (2004), i.e., group-sequential

three-arm clinical trials with a single primary endpoint when assessing the equivalence and efficacy of a generic product.

4.2 Behaviors of the overall power and ASN

Figures 3 (L =2) and 4 (L =5) illustrate the behaviors of overall power, and Figures 5 (L =2) and 6 (L =5) illustrate the behaviors of ASN with correlation under a given sample size per group in the decision-making frameworks for group-sequential strategies with two co-primary endpoints as shown in Tables 2 and 3. The parameter configuration and setting regarding sample sizes and stopping boundaries are the same as other figures.

In the case of L =2, when effect sizes are equal, the powers of all of the decision-making frameworks increase as the correlation goes toward one, but they do not vary with correlation with unequal effect sizes in all the stopping boundary combinations and effect size combinations. The highest power is commonly given by Strategies #4 (DF-A) or/and #1 (DF-B) and the lowest power is commonly given by Strategy #2 (DF-A'). On the other hand, when effect sizes are equal, the ASNs for all of the decision-making frameworks decrease as the correlation goes toward one, but they do not vary with correlation with unequal effect sizes in all the stopping boundary combinations and effect size combinations. The smallest ASN is given by Strategies #2 (DF-B) and #3 (DF-A) and the largest ASN is given by Strategies #1 (DF-A') and #4 (DF-A).

Similar behaviors for the power and ASN are observed in case of L =5. The powers for all of the decision-making frameworks increase as correlation goes toward one, but they do not vary with the correlation with unequal effect sizes in all the stopping boundary combinations and effect size combinations. The highest power is commonly given by Strategies #6 (DF-A) or/and #1 (DF-B) and the lowest power is commonly given by Strategy #2 (DF-A'). On the other hand, when effect sizes are equal, the ASNs for all decision-making frameworks decrease as the correlation goes toward one, but they do not vary with correlation

with unequal effect sizes in all the stopping boundary combinations and effect size combinations. The smallest ASN is given by Strategy #2 (DF-B) and the largest ASN is given by Strategy #6 (DF-A). In summary, delaying the analysis for one of the endpoints increases the power but increases ASN.

5. Summary and discussion

The determination of sample size and the evaluation of power are fundamental and critical elements in the design of a clinical trial. If a sample size is too small then important effects may not be detected, while a sample size that is too large is wasteful of resources and unethically puts more participants at risk than necessary. Recently many clinical trials have been designed with more than one endpoint considered as primary. When utilizing multiple endpoints in clinical trials, we must distinguish between the two inferential goals of clinical trials based on multiple endpoints, i.e., a decision must be made as to whether it is desirable to evaluate the joint effects on *all* endpoints or *at least one* of the endpoints. The former is referred as to "multiple co-primary endpoints" and the latter as to "multiple primary endpoints" (Offen et al., 2007). In this paper, we discuss methods for multiple co-primary endpoints. Co-primary endpoints offer an attractive design feature as they capture a more complete characterization of the effect of an intervention. However co-primary endpoints create challenges in the evaluation of power and the calculation of sample size during trial design as the power is decreased and the sample size is increased with the larger number of endpoints. Currently utilized methods often result in large and impractical sample sizes.

In this paper, as an extension of the work in Asakura et al. (2014, 2015), we consider three decision-making frameworks for group-sequential strategies with multiple co-primary endpoints when appropriately planning for a varying number of analyses for each endpoint and equally or unequally spaced increments of information when the trial is designed to evaluate if a new intervention is superior to a control on *all* of the endpoints. We also consider the use of hierarchical hypothesis testing methodology with the adaptive Type I error

allocation, which was discussed by Tsong et al. (2004). Then we investigate the operating characteristics of group-sequential strategies for clinical trials with multiple co-primary endpoints. Based on the investigations, our findings are summarized in Table 4.

The decision-making framework using hierarchical hypothesis testing with adaptive Type I error allocation (DF-B) has the attractive features of providing higher power and smaller sample sizes compared with the decision-making frameworks with prespecified and fixed Type I error allocation (DF-A or DF-A'). However, the Type I error rate is inflated and depends on the correlation, effect sizes, and the stopping boundary. As seen in clinical trials with two co-primary endpoints, the correlation between the endpoints and the effect size of the first-tested endpoint are the nuisance parameters that determine the stopping boundary and then the level of the Type I error. In practice, use of DF-B should be carefully considered. In a similar but not identical setting, i.e., at least one endpoint with one interim analysis, and one primary and one secondary endpoints, the behavior of the Type I error for hierarchical hypothesis testing has been well-studied (Glimm et al, 2010; Hung et al, 2007; Tamhane et al, 2010). By the analogy between these studies and the investigation given in Appendix 2, one simple solution is to test the hypothesis for the second-tested endpoint only once although further investigation will be required to evaluate more general situations with more than two analyses.

The decision-making framework with prespecified and fixed Type I error allocation (DF-A or DF-A') can adequately control the Type I error rate. They are less powerful than DF-B, but differences in power and required sample sizes are very modest. Especially, when the O'Brien-Fleming-type boundary is selected for both endpoints, there is little difference in power, maximum sample size, and average sample number. DF-A provides the flexibility of selecting differently spaced information levels and different numbers of analyses among the endpoints. In some clinical trials, information for the endpoints may not accrue at the same rate. For example, progression-free survival and overall survival are common endpoints in

oncology trials and require different information times. DF-A is useful when designing clinical trials with such endpoints. Strategic selection regarding the number of analyses with equally or unequally spaced information level among the endpoints may improve the power and reduce the sample sizes. However, when selecting a different number of analyses among the endpoints, early interim evaluations should be carefully evaluated as they can provide higher power but larger average sample numbers. DF-A also offer the option of stopping measurement of an endpoint for which superiority has been demonstrated. This may be desirable if the endpoint is very invasive or expensive. However, these complexities may raise operational challenges. Stopping measurement after interim analysis can raise a major concern about study integrity and can affect the validity of the statistical conclusions reached for a clinical trial. In practice, we should carefully consider how to minimize this risk.

When constructing efficient group-sequential strategies in clinical trials with multiple co-primary endpoints, there are two practical questions. The first question is the choice of the stopping boundary based on an alpha-spending function for each endpoint. If the trial was designed to detect effects on at least one endpoint with a prespecified ordering of endpoints, then the selection of different boundaries for each endpoint (i.e., the O'Brien-Fleming-type for the primary endpoint and the Pocock-type boundary for the secondary endpoint) can provide a higher power than using the same boundary for both endpoints (Glimm et al., 2010; Tamhane et al., 2010). However, as shown in Section 4, the selection of a different boundary has a minimal effect on the overall power and average sample number. In all of the three decision-making frameworks, regardless of equal or unequal effect sizes among the endpoints, the largest power is obtained from the O'Brian-Fleming-type boundary for all of the endpoints, and the lowest is the Pocock-type boundary for all of the endpoints. Regarding the average sample number, the smallest is provided by the Pocock-type boundary for all of the endpoints, the largest is provided by the O'Brian-Fleming-type boundary. One possible scenario for selecting a different boundary is when one endpoint is invasive and stopping to

measurement of the endpoint is desirable as soon as possible, i.e., once the superiority for the endpoint has been demonstrated. Table 5 illustrates the expected number of observations per intervention group for each endpoint based on the decision-making frameworks DF-A under a given maximum sample size in clinical trials with two co-primary endpoints, EP1 and EP2. The expected number of observations for each endpoint is calculated under the hypothetical reference values and provides information on the number of observations anticipated in a group-sequential design in order to reach a decision point for each endpoint. The maximum sample size per intervention group (equally-sized group) is calculated to detect the joint effect for two endpoints (δ_1, δ_2) $(\sigma_1 = \sigma_2 = 1)$ with the overall power of 80% at the significance level of 2.5% for a one-sided test, where one interim and one final analysis are to be performed, the critical values are determined by the O'Brian-Fleming-type boundary, the Pocock-type boundary and their combinations, using the Lan-DeMets alpha-spending method with equally-spaced increments of information, and (δ_1, δ_2) =(0.2, 0.2), (0.2, 0.3) and (0.3, 0.2) are selected. If EP1 is an invasive endpoint, then the combination of the Pocock-type boundary for EP1 and the O'Brian-Fleming-type boundary for EP2, provides the smallest expected number of observations for EP1 in all of the effect size combinations.

Another practical question is the selection of the correlations in the power evaluation and sample size calculation, i.e., whether the observed correlation from external or pilot data should be utilized. One conservative approach is to assume that the correlations are zero even if non-zero correlations are expected. Group-sequential designs discussed in this paper offer the possibility of reducing the sample size compared to fixed sample size designs even if zero correlation is assumed at the design stage. For example, when considering a clinical trial with two co-primary endpoints, 490 participants per intervention group is required to detect a joint effect of equal effect sizes $(\delta_1, \delta_2) = (0.2,\ 0.2)$ with the overall power of 80% at the significance level of 2.5% for a one-sided test in a fixed sample size design, if the correlation between two endpoints is $\rho_{12} = 0.5$. In a group-sequential design with DF-A', if

conservatively assuming zero correlation between the two endpoints, the maximum sample sizes are 518, 523, 528 and 530 corresponding to the number of analyses L =2, 3, 4 and 5, using the O'Brian-Fleming-type boundary based on the Lan-DeMets alpha-spending function for both endpoints with equally-spaced increments of information. Under these maximum sample sizes, the average sample numbers are 488, 455, 442 and 434. The average sample numbers are approximately equal or smaller than the fixed sample designs, depending on the number of analyses. Our experience suggests that when standardized effect sizes are unequal among the endpoints, then the power is not increased with higher correlations. With unequal standardized effect sizes, incorporating the correlation into the sample size calculation at the planning stage may have less of an advantage because the sample size is determined by the smaller effect size.

Acknowledgements Research reported in this publication was supported by JSPS KAKENHI under Grant Number 26330038 and the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Numbers UM1AI104681 and UM1AI068634. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest

The authors have declared no conflict of interest

Appendix 1: Computation time for calculating sample sizes

The program for calculating the power and sample size is coded in FORTAN 77/90, including the subroutine for computing the multivariate normal distribution function values, MVNDST developed by Professor Alan Genz of Washington State University (the subroutine MVNDST is available on at his website http://www.math.wsu.edu/faculty/genz/software/fort77/). The speed of execution heavily depends upon the speed of MVNDST, and is a function of the number of endpoints, the number of analyses, and the required accuracy for computation. In our study, computing the multivariate normal distribution function values using the subroutine

MVNDST began with a maximum number of function values (MAXPTS) of 5000, an absolute error tolerance (ABSEPS) of 0.00001, and a relative error tolerance (RELEPS) of zero. If the estimated absolute error (ERROR) is larger than required tolerance (ABSEPS), i.e., ERROR > ABSEPS, then MAXPTS is increased by 1000 to decrease the estimated absolute error.

To illustrate the computational cost, Table A1 shows the CPU time (in seconds) taken for calculating the sample size on a DELL Precision T7300 (Intel® Xeon® CPU E5-2630/2.60GHz/RAM 8.00GB/32bit operating system) for DF-A with the number of endpoints (k) and the number of analyses (L), where k =2 and 3; L =2, 3, 4, 5, 8 and 10; a common effect size $\delta = \delta_k = 0.2$, and common correlation $\rho = \rho_{kk'} = 0.5$. The sample size is calculated to detect a joint effect on all endpoints with the power of 80% at the significance level of 2.5% for a one-sided test, where the O'Brian-Fleming-type boundary is commonly selected for all of the endpoints, with equally spaced increments of information. We here consider the two methods for calculating sample sizes; one is a grid search to decrease n gradually (decrease by one) until the power under n falls below the desired overall power of $1-\beta$ (Method 1), and the other is an iterative procedure based on linear interpolation to identify n discussed in Hamasaki et al (2013) (Method 2). When the effect sizes are similar among the endpoints and the same stopping boundary is selected for all the endpoints, then the required sample size sample size lies between the two values $n_{\min}$ and $n_{\max}$, where $n_{\min}$ and $n_{\max}$ are the sample sizes calculated to have the power of $1-\beta$, and $(1-\beta)^{1/k}$ for detecting an effect on one endpoint at the significance level of α for a one-sided test, in a group-sequential setting with L analyes. As an initial value for the sample size calculation, $n_{\max}$ is selected for Method 1, and $n_{\min}$ and $n_{\max}$ for Method 2 as Method 2 requires the two initial values.

The table displays the CPU time for Methods 1 and 2. The CPU time increases as the number of analyses increases or as the effect size decreases. Method 1 requires more computing time than Method 2. When calculating sample sizes for a larger number of analyses or when sizing to detect smaller effect sizes, an iterative procedure is required to save the computing time. However, if the number of analysis is larger than 5 and the number of endpoints is larger than 2, even iterative procedures require considerable computing time to compute the sample size. In these situations, a Monte-Carlo simulation-based method provides an alternative although the number of replications for simulations should be carefully chosen to control simulation error in calculating the empirical power.

Appendix 2: Type I error rate in decision-making frameworks of hierarchical hypothesis testing with adaptive Type I error allocation

We discuss the behavior of the Type I error rate in the decision-making framework for clinical trials with co-primary endpoints using hierarchical hypothesis testing with adaptive Type I error allocation (DF-B), discussed in Section 2.3. For illustration, we consider the simplest situation, i.e., a clinical trial with two co-primary endpoints and two analyses are planned. The probability for rejecting the null hypothesis for DF-B is given by

$$\Pr[Z_{11} > c_{11}(2), Z_{21} > c_{21}(2) | \delta_1, \delta_2, \rho]$$
$$+\Pr[Z_{11} > c_{11}(2), Z_{21} \leq c_{21}(2), Z_{22} > c_{22}(2) | \delta_1, \delta_2, \rho]$$
$$+\Pr[Z_{11} \leq c_{11}(2), Z_{12} > c_{12}(2), Z_{21} > c_{21}(1) | \delta_1, \delta_2, \rho],$$

where $c_{kl}(L_k)$ are the critical values for kth endpoint at the lth analysis with the maximum number of analyses L_k. From the definitions, it is clear that the Type I error rate is a function of two nuisance parameters, i.e., correlation and effect sizes. The critical values for Endpoint 2 in the first two terms are the same as those seen in the prefixed Type I error allocation although they are determined by the result of Endpoint 1. However the critical value for Endpoint 2 in the third term clearly depends on the result of Endpoint 1.

As seen in Figures 5 and 6, the Type I error rate is inflated, i.e., higher than the targeted significance level when $(\delta_1, \delta_2) = (0.2, 0.0)$. Therefore, to evaluate the Type I error, we consider just the situation of $\delta_1 \neq 0$ and $\delta_2 = 0$. When $\rho = 0$, the Type I error rate for DF-B is

$$\begin{aligned}\alpha_B &= \Pr[Z_{11} > c_{11}(2) | \delta_1 \neq 0] \\ &\quad \times (\Pr[Z_{21} > c_{21}(2) | \delta_2 = 0] + \Pr[Z_{21} \leq c_{21}(2), Z_{22} > c_{22}(2) | \delta_2 = 0]) \\ &\quad + \Pr[Z_{11} \leq c_{11}(2), Z_{12} > c_{12}(2) | \delta_1 \neq 0] \Pr[Z_{21} > c_{21}(1) | \delta_2 = 0] \\ &\quad \leq \Pr[Z_{11} > c_{11}(2) | \delta_1 \neq 0]\alpha + \Pr[Z_{11} \leq c_{11}(2), Z_{12} > c_{12}(2) | \delta_1 \neq 0]\alpha \\ &\quad = (1 - \beta_1)\alpha,\end{aligned}$$

where $1 - \beta_1$ is the power for detecting the effect size for Endpoint 1. So that when $\rho = 0$, the Type I error rate for DF-B is not larger than the targeted significance level. However, when $\rho > 0$, the Type I error rate for DF-B is inflated, depending on δ_1 and ρ. To illustrate how the Type I error rate for DF-B changes with δ_1 and ρ, Figures A1 to A3 provide the behaviors of the overall Type I error rate for DF-B as a function of correlation (ρ) and effect size for Endpoint 1 (δ_1) under a given sample size per group (equally-sized groups) in group-sequential strategies for clinical trials with two co-primary endpoints and two analyses. Also the four stopping-boundary combinations are considered as the critical value for Endpoint 2 depends on the effect size for Endpoint 1; (i) the OF for both endpoints (OF-OF), (ii) the OF for δ_1 and the PC for δ_2 (OF-PC), (iii) the PC for δ_1 and the OF for δ_2 (PC-OF), and (iv) the PC for both endpoints (PC-PC). The sample sizes per group are calculated 86 for Figure A1, 516 for Figure A2, 2068 for Figure A3 to detect the joint effect of $(\delta_1, \delta_2) = (0.5, 0.5)$, (0.2,0.2), and (0.1, 0.1), with the power of 80% at the significance level of 2.5% for a one-sided test in a fixed sample size design. For the assessment of the Type I error rate, the effect size combination $(\delta_1, \delta_2) = (\delta_1^*, 0)$ is considered, where $0 \leq \delta_1^* \leq 1$. Figures A1 to A3 show that the Type I error rate for DF-B is inflated with higher correlation and smaller effect size

for Endpoint 1, especially with smaller sample sizes, and PC-PC and OF-PC stopping-boundary combinations. The third term of the Type I error rate for DF-B, $\Pr[Z_{11} \leq c_{11}(2), Z_{12} > c_{12}(2), Z_{21} > c_{21}(1) | \delta_1, \delta_2, \rho]$ is relevant for adaptive Type I error allocation as the critical value for Endpoint 2 is determined based on the result on the Endpoint 1 and contributes to inflation of the Type I error rate.

References

Asakura, K., Hamasaki, T., Evans, S.R., Sugimoto, T., and Sozu, T. (2015),"Sample Size Determination in Group-Sequential Clinical Trials with Two Co-Primary Endpoints," in Applied Statistics in Biomedicine and Clinical Trial Design, by Z. Chen et al. (eds.), Springer (in press)

Asakura, K., Hamasaki, T., Sugimoto, T., Hayashi, K., Evans, S.R., and Sozu, T. (2014), "Sample Size Determination in Group-Sequential Clinical Trials with Two Co-Primary Endpoints," Statistics in Medicine, 33, 2897–2913. DOI: 10.1002/sim.6154

Cheng, Y., Ray, S., Chang, M., and Menon, S. (2014), "Statistical Monitoring of Clinical Trials with Multiple Co-Primary Endpoints Using Multivariate B-value," Statistics in Biopharmaceutical Research, 6, 241-250. DOI: 10.1080/19466315.2014.923324

Committee for Medicinal Products for Human Use (CHMP) (2008), "Guideline on Medicinal Products for the Treatment Alzheimer's Disease and Other Dementias" (CPMP/EWP/553/95 Rev.1). EMEA: London, 2008.

Cui, L., Hung, H.M.J., and Wang, S.J. (1999), "Modification of Sample Size in Group Sequential Clinical Trials", Biometrics, 55, 853–857. DOI: 10.1111/j.0006-341X.1999.00853.x.

Chuang-Stein, C., Stryszak, P., Dmitrienko, A., and Offen, W. (2007), "Challenge of Multiple Co-Primary Endpoints: A New Approach," Statistics in Medicine, 26, 1181-1192. DOI: 10.1002/sim.2604.

Eaton, M.L., and Muirhead, R.J. (2007), "On Multiple Endpoints Testing Problem," Journal of Statistical Planning & Inference, 137, 3416-3429. DOI: 10.1016/j.jspi.2007.03.021.

Glimm, E., Maurer, W., and Bretz, F. (2010), "Hierarchical Testing of Multiple Endpoints in Group-Sequential Trials," Statistics in Medicine, 29, 219–228. DOI: 10.1002/sim.3748

Hamasaki, T., Sugimoto, T., Evans, S.R., and Sozu, T. (2013), "Sample Size Determination for Clinical Trials with Co-Primary Outcomes: Exponential Event Times," Pharmaceutical Statistics, 12, 28-34. DOI: 10.1002/pst.1545.

Hung, H.M.J., and Wang, S.J. (2009), "Some Controversial Multiple Testing Problems in Regulatory Applications," Journal of Biopharmaceutical Statistics, 19, 1-11. DOI: 10.1080/10543400802541693.

Hung, H.M.J., Wang, S.J., and O'Neill, R.T. (2007), "Statistical Considerations for Testing Multiple Endpoints in Group Sequential or Adaptive Clinical Trials," Journal of Biopharmaceutical Statistics, 17, 1201-1210. DOI: 10.1080/10543400701645405.

Genz, A. (1992), "Numerical Computation of Multivariate Normal Probabilities," Journal of Computational and Graphical Statistics, 1, 141-149.

Julious, S., and Mclntyre, N.E. (2012), "Sample Sizes for Trials Involving Multiple Correlated Must-Win Comparisons," Pharmaceutical Statistics, 11, 177-185. DOI: 10.1002/pst.515.

Kordzakhia, G., Siddiqui, O., and Huque, M.F. (2010), "Method of Balanced Adjustment in Testing Co-Primary Endpoints," Statistics in Medicine, 29, 2055-2066. DOI: 10.1002/sim.3950.

Lan, K.K.G., and DeMets, D.L. (1983), "Discrete Sequential Boundaries for Clinical Trials," Biometrika, 70, 659-663. DOI: 10.1093/biomet/70.3.659

O'Brien, P.C., and Fleming, T.R. (1979), "A Multiple Testing Procedure for Clinical Trials", Biometrics, 35, 549-556. DOI: 10.2307/2530245

Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson, J.D., Krishen, A., Liu, T., Ryder, S., Sankoh, A.J., Wang, J., and Yeh, C.H. (2007), "Multiple Co-Primary Endpoints: Medical and Statistical Solutions," Drug Information Journal, 41, 31-46. DOI: 10.1177/009286150704100105.

Pocock, S.J. (1977), "Group Sequential Methods in the Design and Analysis of Clinical Trials," Biometrika, 64, 191–199. DOI: 10.1093/biomet/64.2.191

Senn, S., and Bretz, F. (2007), "Power and Sample Size when Multiple Endpoints Are Considered," Pharmaceutical Statistics, 6, 161-170. DOI: 10.1002/pst.301.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2010), "Sample Size Determination in Clinical Trials with Multiple Co-Primary Binary Endpoints," Statistics in Medicine, 29, 2169-2179. DOI: 10.1002/sim.3972

Sozu, T., Sugimoto, T., and Hamasaki, T. (2011), "Sample Size Determination in Superiority Clinical Trials with Multiple Co-Primary Correlated Endpoints," Journal of Biopharmaceutical Statistics, 21, 1-19. DOI: 10.1080/10543406.2011.551329.

Sozu, T., Sugimoto, T., and Hamasaki, T. (2012), "Sample Size Determination in Clinical Trials with Multiple Co-Primary Endpoints Including Mixed Continuous and Binary Variables," Biometrical Journal, 54, 716-29. Doi: 10.1002/bimj.201100221.

Sozu, T., Sugimoto, T., Hamasaki, T., and Evans S.R. (2015), "Sample Size Determination in Clinical Trials with Multiple Primary Endpoints," Springer (in press).

Sugimoto, T., Sozu, T., and Hamasaki, T. (2012), "A Convenient Formula for Sample Size Calculations in Clinical Trials with Multiple Co-Primary Continuous Endpoints," Pharmaceutical Statistics, 11, 118-128. DOI: 10.1002/pst.505.

Sugimoto, T., Sozu, T., Hamasaki, T., and Evans, S.R. (2013), "A Logrank Test-Based Method for Sizing Clinical Trials with Two Co-Primary Time-to-Event Endpoints," Biostatistics, 14, 409-421. DOI: 10.1093/biostatistics/kxs057.

Xiong, C., Yu, K., Gao, F., Yan, Y., and Zhang, Z. (2005), "Power and Sample Size for Clinical Trials When Efficacy Is Required in Multiple Endpoints: Application to An Alzheimer's Treatment Trial," Clinical Trials, 2, 387-393. DOI: 10.1191/1740774505cn112oa.

Tamhane, A.C., Mehta, C.R., and Liu, L. (2010). "Testing A Primary and Secondary Endpoint in A Group Sequential Design," Biometrics, 66, 1174–1184. DOI: 10.1111/j.1541-0420.2010.01402.x

Tsong, Y., Zhang, J., and Wang, S.J. (2004), "Group Sequential Design and Analysis of Clinical Equivalence Assessment for Generic Nonsystematic Drug Products," Journal of Biopharmaceutical Statistics, 14, 359–373. DOI: 10.1081/BIP-120037186.

Table 1. O'Brien-Fleming-type boundary corresponding to the rejection of the null hypothesis for the first- and second-tested endpoints in hierarchical hypothesis testing with the adaptive Type I error allocation

Interim analysis and Information time for the second-tested endpoint		Interim analysis and information time for the first-tested endpoint				
		1st (0.2)	2nd (0.4)	3rd (0.6)	4th (0.8)	Final (1.0)
		4.8769	3.3569	2.6803	**2.2898**	2.0310
1st	(0.2)	4.8769	3.3569	2.6803	2.2898	2.0310
2nd	(0.4)		3.3569	2.6802	2.2898	2.0310
3rd	(0.6)			2.6686	2.2887	2.0306
4th	(0.8)				**2.2504**	**2.0249**
Final	(1.0)					1.9600

Table 2. Several group-sequential strategies for clinical trials with two endpoints: Two planned analyses for Endpoint 1

Strategy No.	Decision-making framework	Number of analyses for each endpoint		Information time 1/2	Information time 1
1	DF-B	EP1	2	○	○
		EP2	2	◌	◌
2	DF-A'	EP1	2	○	○
		EP2	2	○	○
3	DF-A	EP1	2	○	○
		EP2	2	○	○
4	DF-A	EP1	2	○	○
		EP2	1		○

○: Endpoint is tested at the information time
◌: If superiority has been established for the Endpoint 1 (EP1), then the second endpoint (EP2) is tested.

Table 3. Several group-sequential strategies for clinical trials with two endpoints: Five planned analyses for Endpoint 1

Strategy No.	Decision-making framework	Number of analyses for each endpoint		Information time 1/5	2/5	3/5	4/5	1
1	DF-B	EP1	5	○	○	○	○	○
		EP2	5	◌	◌	◌	◌	◌
2	DF-A'	EP1	5	○	○	○	○	○
		EP2	5	○	○	○	○	○
3	DF-A	EP1	5	○	○	○	○	○
		EP2	5	○	○	○	○	○
4	DF-A	EP1	5	○	○	○	○	○
		EP2	4		○	○	○	○
5	DF-A	EP1	5	○	○	○	○	○
		EP2	3			○	○	○
6	DF-A	EP1	5	○	○	○	○	○
		EP2	2				○	○

○: Endpoint is tested at the information time
◌: If superiority has been established for the Endpoint 1 (EP1), then the second endpoint (EP2) is tested.

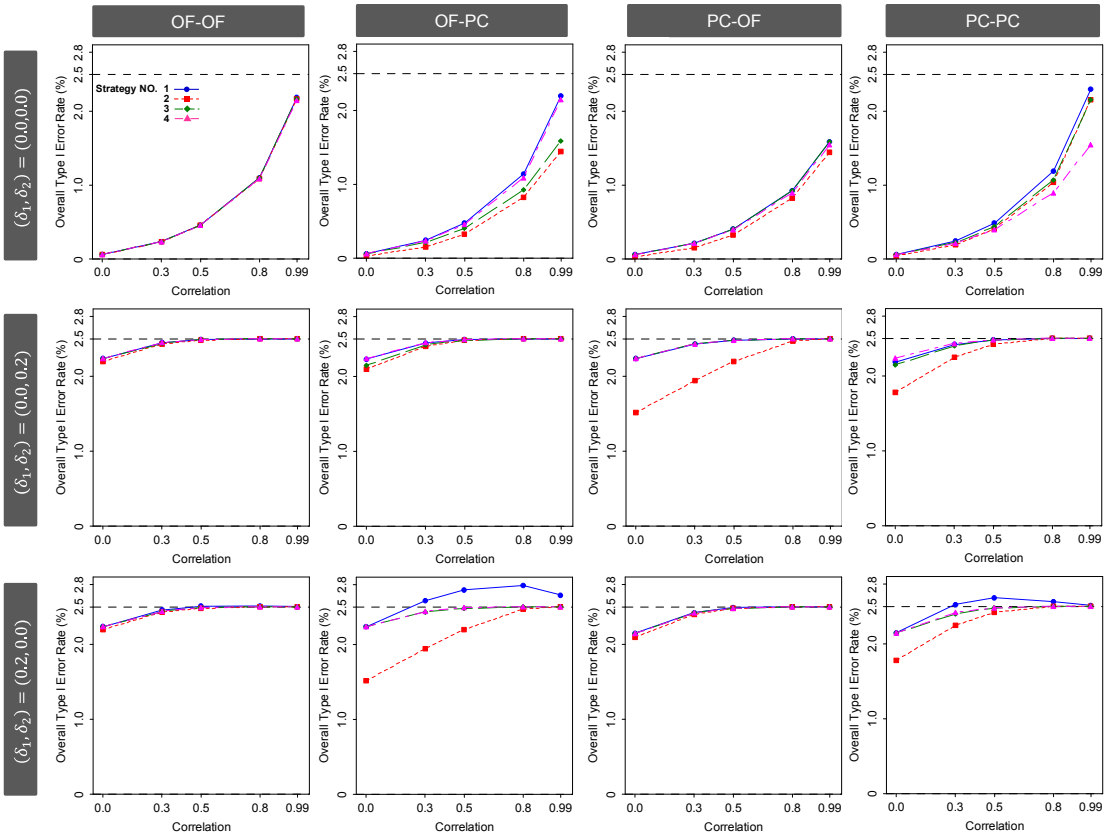

Figure 1. Behavior of the overall Type I error rate under a given maximum sample size per group (equally-sized groups) in group-sequential strategies for clinical trials with two co-primary endpoints as shown in Table 1 (L =2)

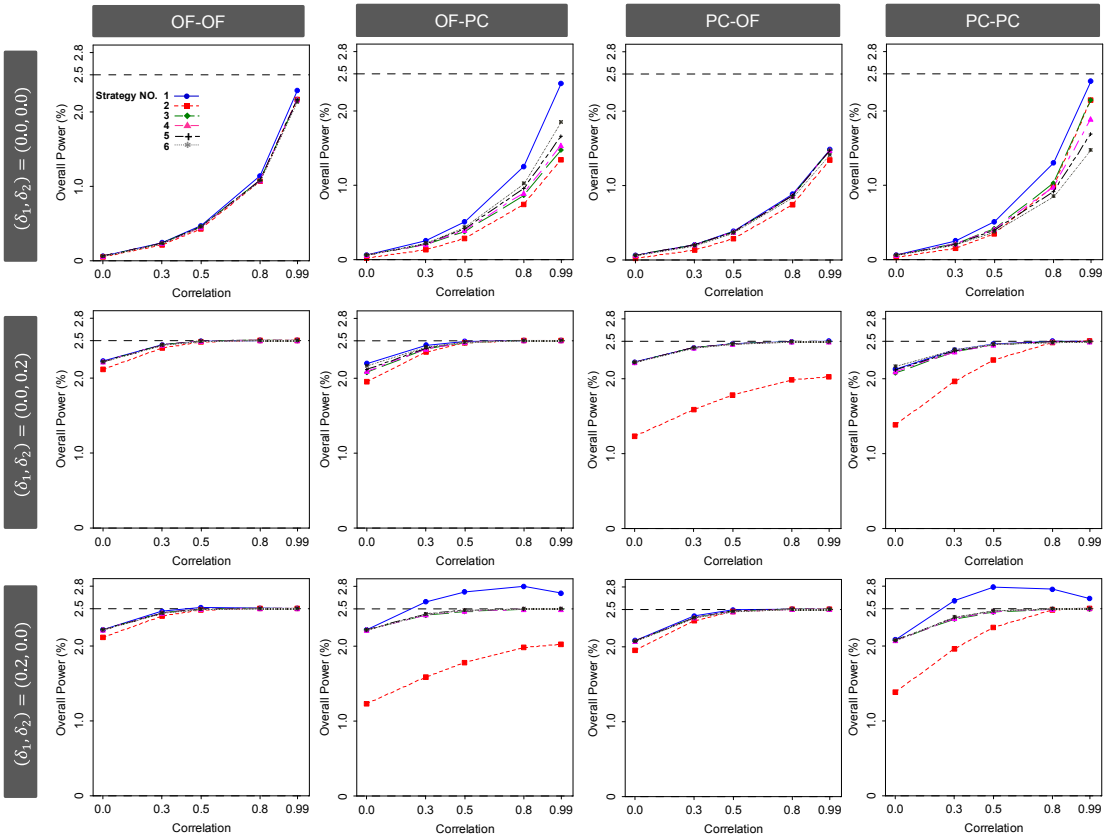

Figure 2. Behavior of the overall Type I error rate under a given maximum sample size per group (equally-sized groups) in group-sequential strategies for clinical trials with two co-primary endpoints as shown in Table 2 ($L = 5$)

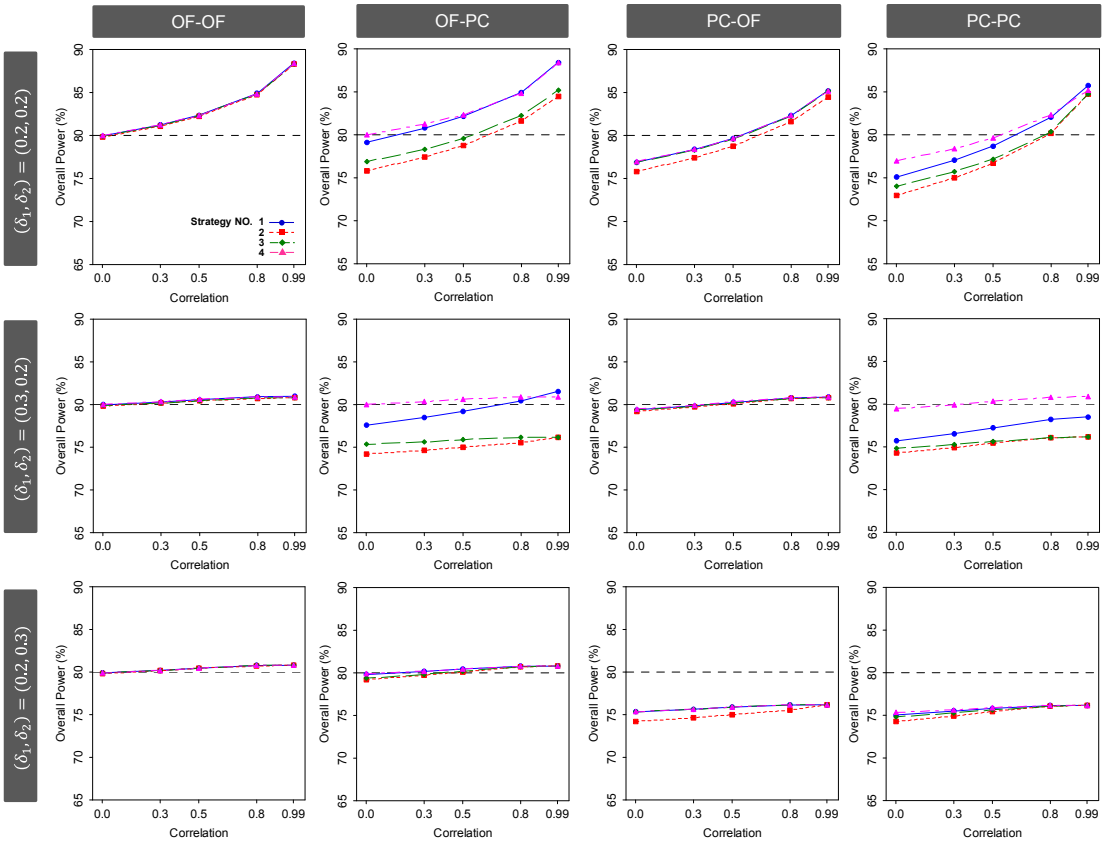

Figure 3. Behavior of the overall power under a given sample size per group (equally-sized groups) in group-sequential strategies for clinical trials with two co-primary endpoints as shown in Table 1 (L =2)

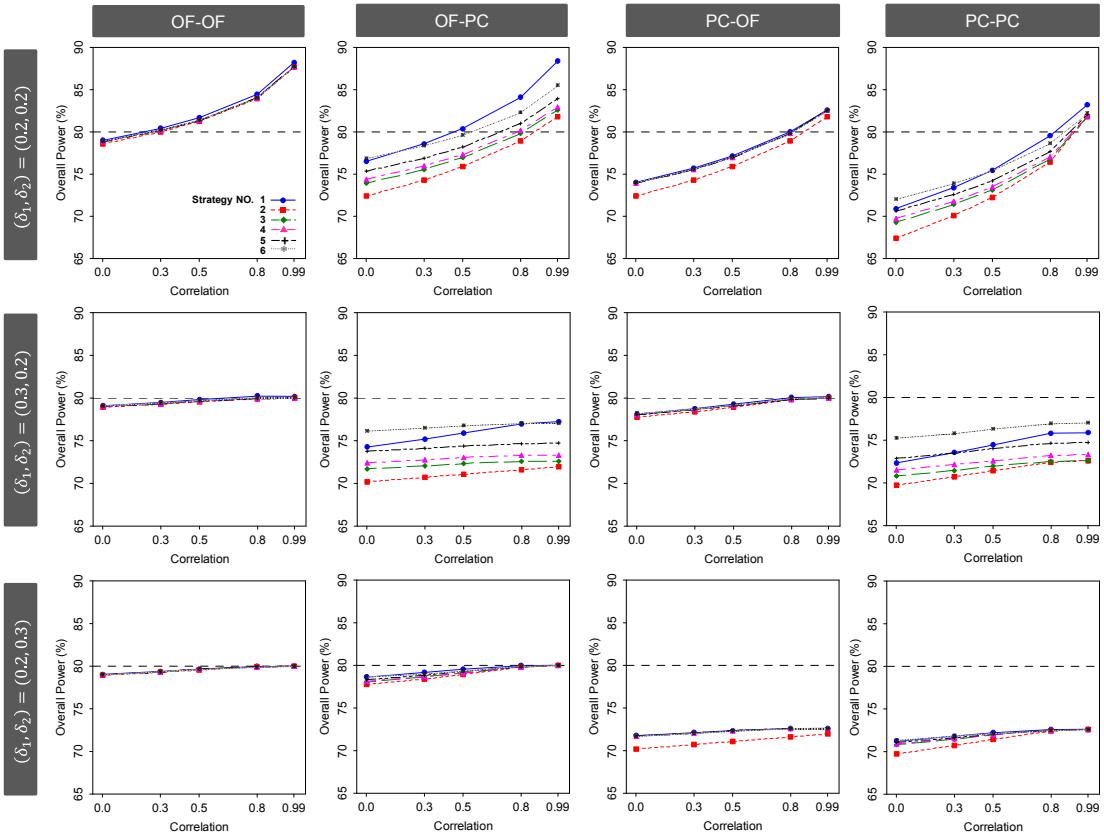

Figure 4. Behavior of the overall power under a given sample size per group (equally-sized groups) in group-sequential strategies for clinical trials with two co-primary endpoints as shown in Table 2 (L =5)

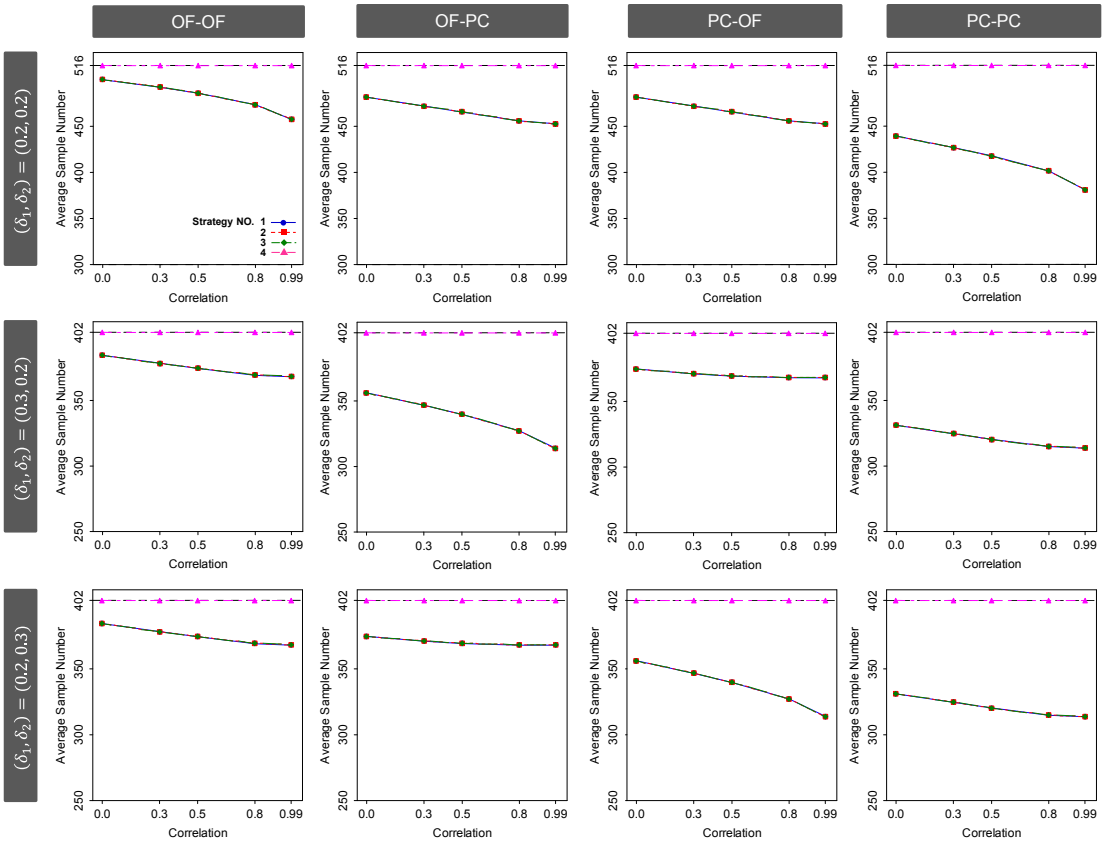

Figure 5. Behavior of the ASN under a given sample size per group (equally-sized groups) in group-sequential strategies for clinical trials with two co-primary endpoints as shown in Table 1 (L =2)

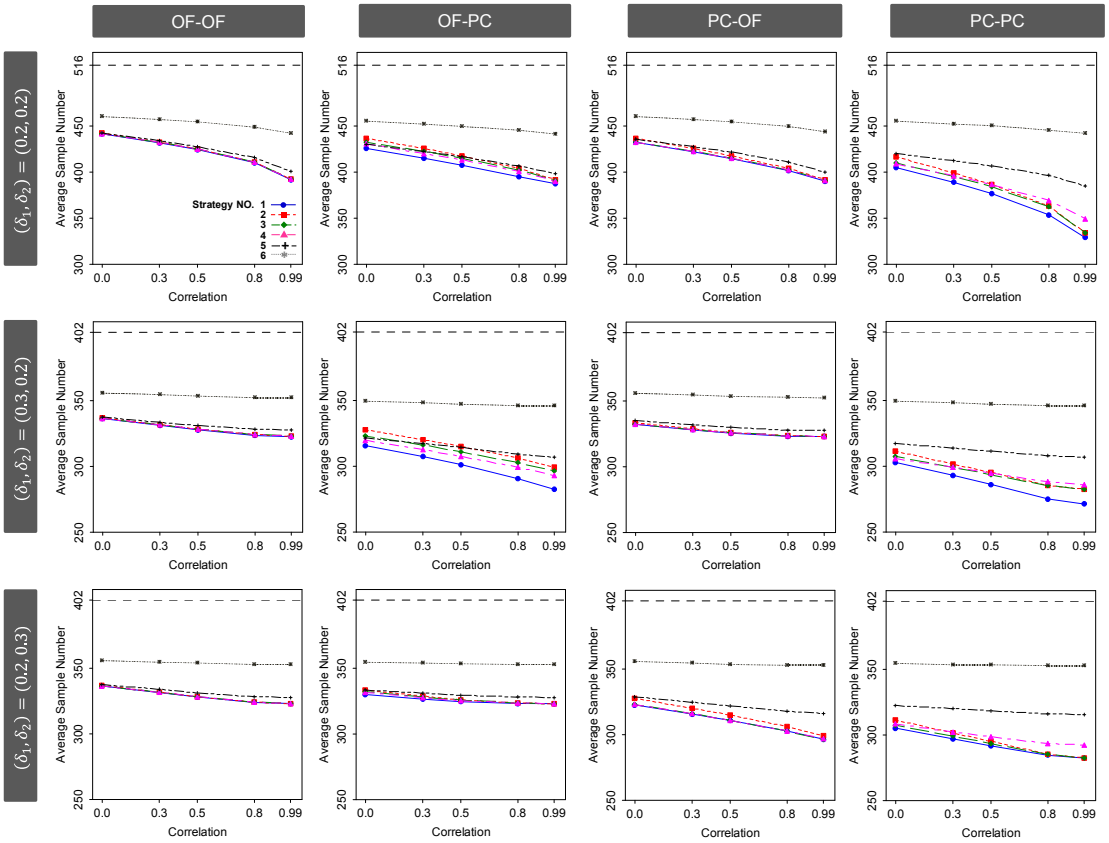

Figure 6. Behavior of the ASN under a given sample size per group (equally-sized groups) in group-sequential strategies for clinical trials with two co-primary endpoints as shown in Table 2 (L =5)

Table 4. Advantages and disadvantages of the three decision-making frameworks in clinical trials with multiple co-primary endpoints

Decision-making framework	Advantages	Disadvantages
DF-A	• Controls the Type I error rate adequately • Flexible to allow the option of selecting different timings for interim looks among the endpoints- this is useful when designing clinical trials with the endpoints requiring different information times such as progression-free survival and overall survival • Possible to stop measuring an endpoint for which superiority has been demonstrated – this is desirable if the endpoint is very invasive or expensive (e.g., data from a liver biopsy or gastro-fiberscope, or data from expensive imaging	• Conservative as the rejection region of the null hypothesis is restricted with the number of endpoints • Difficult to maintain the integrity and validity of clinical trial if stop measuring an endpoint for which superiority has been demonstrated.
DF-A’	• Controls the Type I error rate adequately • Makes the decision-making simple and easy to use practice	• Conservative as the rejection region of the null hypothesis is restricted with the number of endpoints • Cannot stop measuring an endpoint for which superiority has been demonstrated • Provides the lowest power and largest sample sizes among the decision-frameworks
DF-B	• Provides a slightly higher power and then smaller sample sizes compared with the decision-making frameworks with prefixed Type I error allocation (DF-A or DF-A’)	• Needs to prespecify the order of hypothesis testing for each endpoint even the endpoints are equally important • Inflates the Type I error rate, depending on the correlation among the endpoints, effect sizes, and stopping boundary based on the alpha-spending function

Table 5. The expected number of observations per intervention group for each endpoint based on the decision-making framework DF-A under a given maximum sample size in clinical trials with two co-primary endpoints, EP1 and EP2. The maximum sample size per intervention group (equally-sized groups) is calculated to detect the joint effect for two endpoints (δ_1, δ_2) $(\sigma_1 = \sigma_2 = 1)$ with the overall power of 80% at the significance level of 2.5% for a one-sided test, where one interim and one final analysis are to be performed, the critical values are determined by the O'Brian-Fleming-type boundary, the Pocock-type boundary or their combinations, using the Lan-DeMets alpha-spending method with equally-spaced increments of information, and (δ_1, δ_2) =(0.2, 0.2), (0.2, 0.3) and (0.3, 0.2) are selected.

Effect size (δ_1, δ_2)	Expected # of observations and sample sizes		Stopping boundary combinations			
			OF-OF	OF-PC	PC-OF	PC-PC
(0.2, 0.2)	Expected # of observations	EP1	454.2	474.1	390.5	403.4
		EP2	454.2	390.5	474.1	403.4
	Maximum sample size		518	547	547	574
	Average Sample number		502.3	505.3	505.3	472.6
(0.2, 0.3)	Expected # of observations	EP1	368.8	372.7	338.5	340.7
		EP2	298.3	243.0	316.4	259.4
	Maximum sample size		403	408	446	450
	Average Sample number		385.2	379.5	383.5	357.4
(0.3, 0.2)	Expected # of observations	EP1	298.3	316.4	243.0	259.4
		EP2	368.8	338.5	372.7	340.7
	Maximum sample size		403	446	408	450
	Average Sample number		385.2	383.5	379.5	357.4

Table A1. CPU Time (seconds) for computing the sample size and calculated sample size (n_L^*). The sample size is calculated by two methods to detect a joint effect on all endpoint with the power of 80% at the significance level of 2.5% for a one-sided test, where the O'Brian-Fleming-type boundary is selected for the two endpoints, with equally spaced increments of information

# of endpoints	# of analyses	Method 1		Method 2	
		CPU time(sec)	(n_L^*)	CPU time (sec)	(n_L^*)
2	2	9.1	(492)	1.4	(492)
	3	33.6	(497)	5.1	(497)
	4	36.8	(501)	7.9	(501)
	5	185.8	(503)	27.5	(503)
	8	12669.2	(508)	1946.5	(508)
	10	147315.2	(510)	23007.9	(510)
3	2	91.2	(547)	11.7	(547)
	3	245.0	(552)	24.9	(552)
	4	1853.9	(557)	234.4	(557)
	5	16192.5	(560)	2017.2	(560)
	8	>1000000.0	–	937133.2	(565)
	10	>1000000.0	–	>1000000.0	–

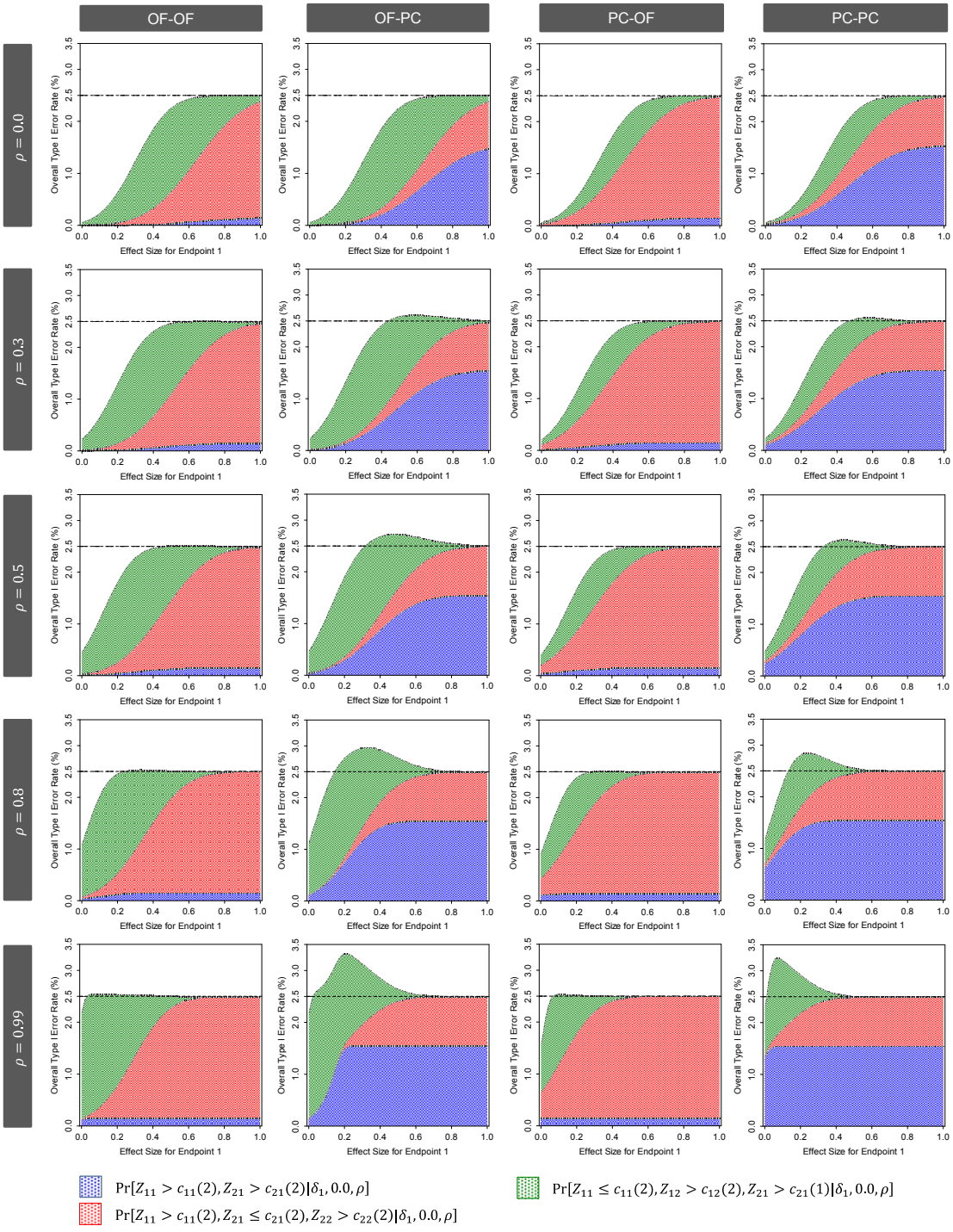

Figure A1. Behavior of the overall Type I error rate for DF-B as a function of correlation (ρ) and effect size for Endpoint 1 (δ_1)-under a given sample size per group (equally-sized groups: $r = 1$) in group-sequential strategies for clinical trials with two co-primary endpoints and two analyses, where $\sigma_1 = \sigma_2 = 1.0$. For the assessment of the Type I error rate, $\delta_2 = 0.0$ is assumed. The sample size per group of 86 is calculated to detect the joint effect of (δ_1, δ_2) =(0.5, 0.5) with the power of 80% at the significance level of 2.5% for a one-sided test in a fixed sample size design

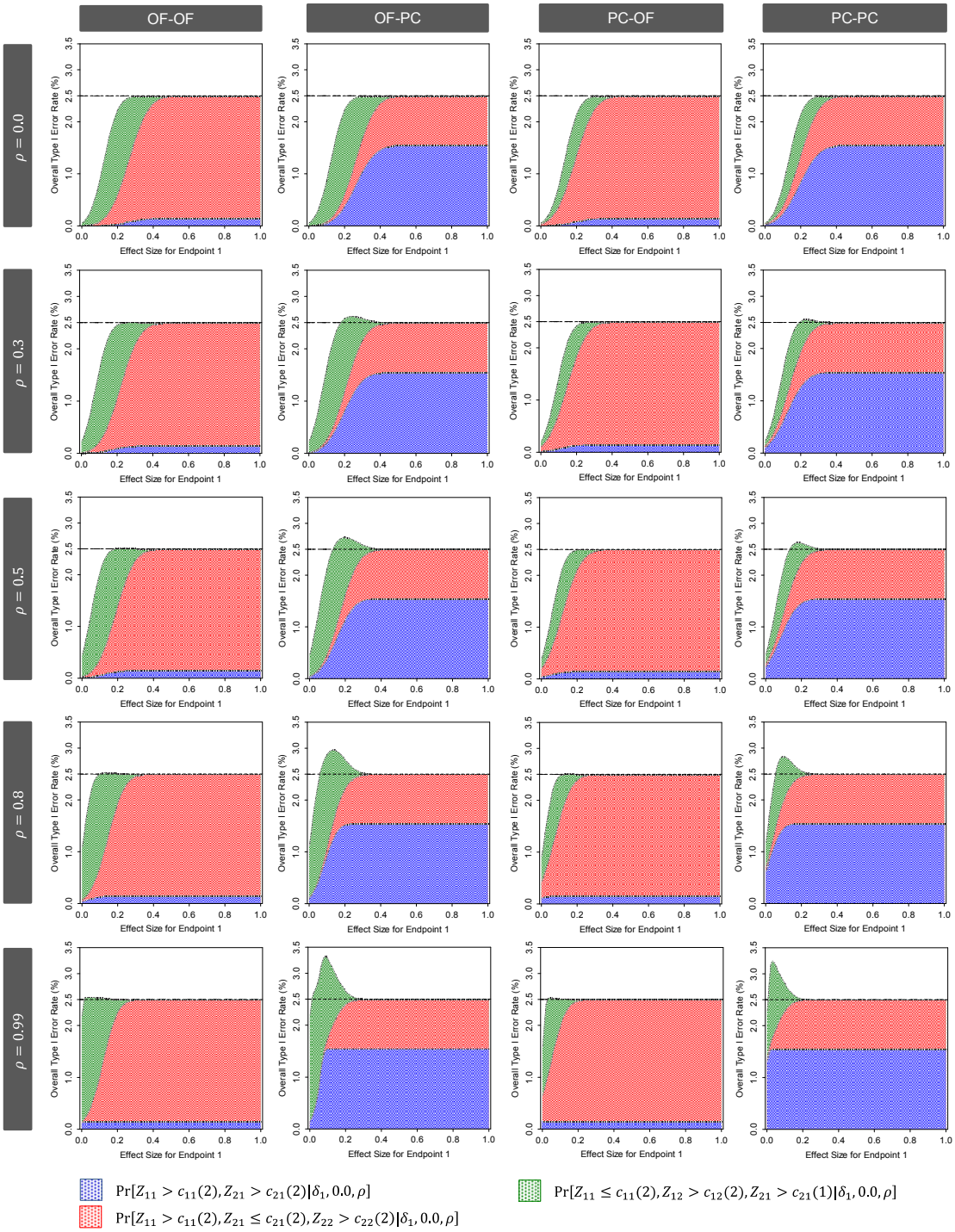

Figure A2. Behavior of the overall Type I error rate for DF-B as a function of correlation (ρ) and effect size for Endpoint 1 (δ_1) under a given sample size per group (equally-sized groups: $r = 1$) in group-sequential strategies for clinical trials with two co-primary endpoints and two analyses, where $\sigma_1 = \sigma_2 = 1.0$. For the assessment of the Type I error rate, $\delta_2 = 0.0$ is assumed. The sample size per group of 517 is calculated to detect the joint effect of (δ_1, δ_2) =(0.2, 0.2) with the power of 80% at the significance level of 2.5% for a one-sided test in a fixed sample size design

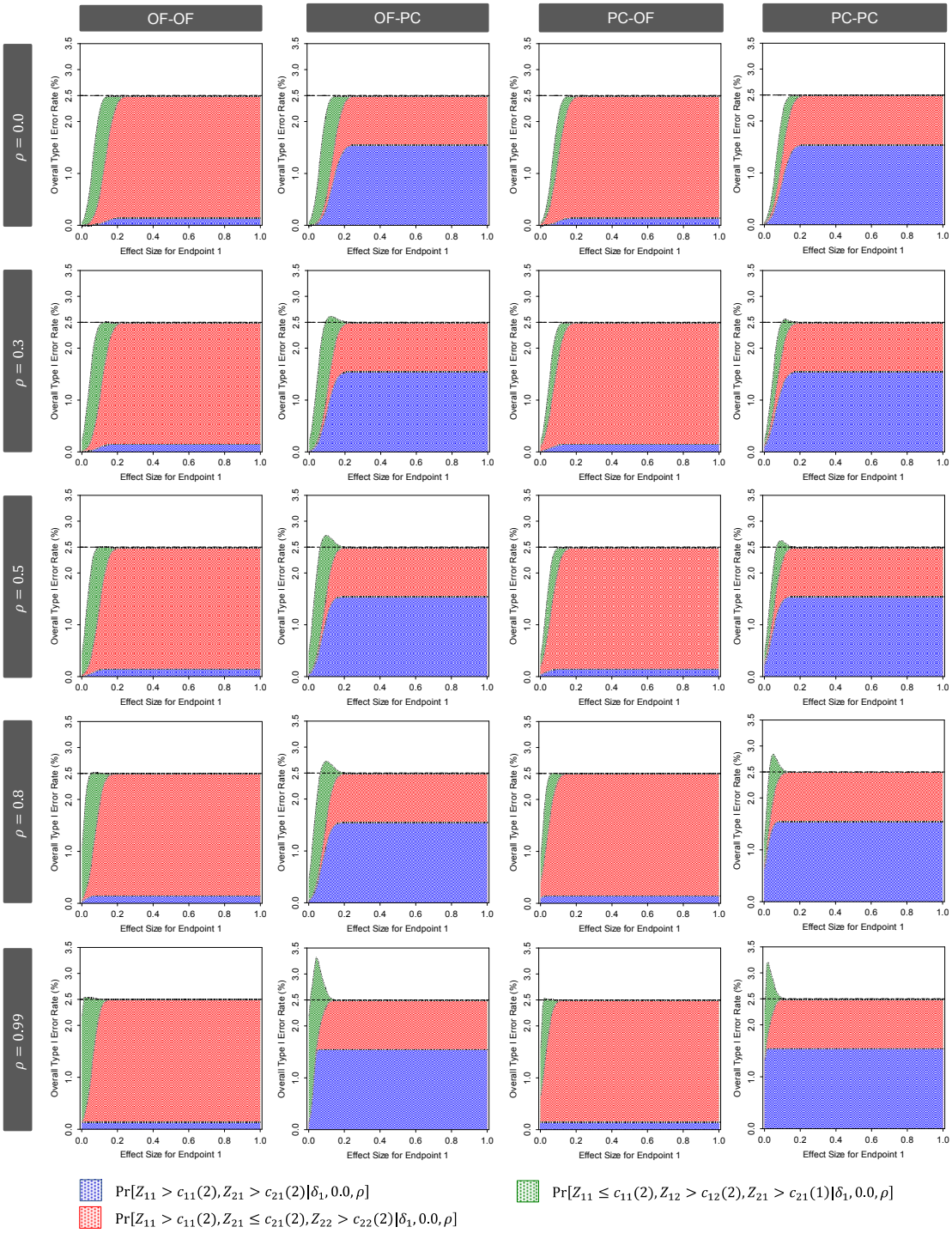

Figure A3. Behavior of the overall Type I error rate for DF-B as a function of correlation (ρ) and effect size for Endpoint 1 (δ_1) under a given sample size per group (equally-sized groups: $r = 1$) in group-sequential strategies for clinical trials with two co-primary endpoints and two analyses, where $\sigma_1 = \sigma_2 = 1.0$. For the assessment of the Type I error rate, $\delta_2 = 0.0$ is assumed. The sample size per group of 2,068 is calculated to detect the joint effect of (δ_1, δ_2) =(0.1, 0.1) with the power of 80% at the significance level of 2.5% for a one-sided test in a fixed sample size design